AD ____________________

Award Number: **DAMD17-96-1-6229**

TITLE: **Cancers Missed on Mammography**

PRINCIPAL INVESTIGATOR: **Robert M. Nishikawa, Ph.D.**

CONTRACTING ORGANIZATION: **University of Chicago**
Chicago, Illinois 60637

REPORT DATE: **October 1999**

TYPE OF REPORT: **Annual**

PREPARED FOR: **U.S. Army Medical Research and Materiel Command**
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT: **Approved for public release; Distribution Unlimited**

REPORT DOCUMENTATION PAGE

Form Approved OMB No. 074-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	October 1999	Annual (23-Sep-98 - 22-Sep-99)

4. TITLE AND SUBTITLE
Cancers Missed on Mammography

5. FUNDING NUMBERS
DAMD17-96-1-6229

6. AUTHOR(S)
Robert M. Nishikawa, Ph.D.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
University of Chicago
Chicago, Illinois 60637

E-MAIL:
r-nishikawa@uchicago.edu

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release; Distribution Unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 Words)*

Observer error in reading screening mammograms has been identified as a significant factor in delayed diagnosis of breast cancer. The magnitude of the problem is estimated to be about 30% of potentially detectable cancers are overlooked for one or more years before detection. Computer-aided diagnosis (CAD) programs have been developed to aid radiologists in the detection task, and pre-clinical studies have shown that CAD applied to digitized mammography films can flag about 50% of radiologists' observational oversights. Preliminary study has also shown a wide variability in radiologist observer performance. The purpose of this investigation is to test how many additional cancers are detected by radiologists using CAD, in an observer study using an enriched mixture of cancers (400 cases, with benign to malignant ratio 3:1). Based on previous work, the expectation is that average cancer detection will improve about 15% with the use of CAD, similar to what has been shown when two human observers (double reading) read the same cases. The phases of the project include collection of case material [missed cancers and normals], characterization and digitization of the films, running the CAD programs, performance of the observer study, and analysis of the results.

14. SUBJECT TERMS
Mammography, computer-aided diagnosis, missed cancers

15. NUMBER OF PAGES
17

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unlimited

NSN 7540-01-280-5500

Standard Form 298 (Rev. 2-89)
Prescribed by ANSI Std. Z39-18
298-102

FOREWORD

Opinions, interpretations, conclusions and recommendations are those of the author and are not necessarily endorsed by the U.S. Army.

___ Where copyrighted material is quoted, permission has been obtained to use such material.

___ Where material from documents designated for limited distribution is quoted, permission has been obtained to use the material.

___ Citations of commercial organizations and trade names in this report do not constitute an official Department of Army endorsement or approval of the products or services of these organizations.

N/A In conducting research using animals, the investigator(s) adhered to the "Guide for the Care and Use of Laboratory Animals," prepared by the Committee on Care and use of Laboratory Animals of the Institute of Laboratory Resources, national Research Council (NIH Publication No. 86-23, Revised 1985).

KM For the protection of human subjects, the investigator(s) adhered to policies of applicable Federal Law 45 CFR 46.

X In conducting research utilizing recombinant DNA technology, the investigator(s) adhered to current guidelines promulgated by the National Institutes of Health.

X In the conduct of research utilizing recombinant DNA, the investigator(s) adhered to the NIH Guidelines for Research Involving Recombinant DNA Molecules.

N/A In the conduct of research involving hazardous organisms, the investigator(s) adhered to the CDC-NIH Guide for Biosafety in Microbiological and Biomedical Laboratories.

Kurt Mueller 2/17/2000

PI - Signature Date

4. TABLE OF CONTENTS

5. INTRODUCTION

Double reading of mammograms has been shown to significantly increase the number of cancers detected.[1-5] Computer-aided diagnosis (CAD) has been proposed as an efficient method of implementing double reading.[6] For CAD to be effective computers must find cancers that are missed by radiologists, and radiologists must react appropriately to the computer prompts. Others and we have found that computer detection schemes can find over 50% of the observational misses made by radiologists reading mammograms.[7-9] Our current study is designed to show that CAD can help detect cancers that they might otherwise overlook. We will collect a large database of cancers already missed by radiologists in routine clinical practice, and will test observers without and with the aid of CAD. It is expected that radiologists will detect about 10 to 15% more cancers using CAD, which would have important implications for bringing this technique into clinical practice. We will also learn much more about the reasons for and types of radiologist misses on mammography.

6. BODY OF REPORT

6.1 Tasks

There are five tasks in the Statement of Work, which are listed below.

Task 1. Preparation of review forms and finalization of eligibility characteristics for cases to be entered into the missed lesion database.

Task 2. Accumulation of database cases and copying/digitizing 100 missed malignant cases and 300 normal cases, with categorization of features and characteristics of the malignant case. Verification of missed lesion cases. Ongoing data entry.

Task 3. Computer runs producing hard copy of computer output for use in observer experiment and preparation of cases for observer experiment. Ongoing data entry of computer accuracy and truth table for missed lesion database. Final design of details of observer performance study.

Task 4. An observer experiment conducted on 15 observers at about 3 hours per session, with 6 sessions per observer spaced at 2-3 months apart. Goal is to perform 2 observation sessions and analysis minimum per week, entering observation data into a computer database. Ongoing data entry.

Task 5. Final analysis of data comparing CAD observer results with non-CAD results and observer variability, and preparation of report summarizing the results of the observer experiment and the clinical characteristics of the missed lesions.

6.1.1 Preparation of forms

A copy of the review form is attached. The eligibility criteria are as follows:

1. Patients who have had screen-film mammograms read at the participating mammography facilities.

2. For cases of missed lesions, the mammogram had to be read clinically as normal in the area where a cancer subsequently developed, and the error had to be one of observation (failure to see the lesion) rather than interpretation (seeing the lesion and categorizing it as benign). In cases where the cancer is visible on multiple examinations prior to diagnosis, the two expert mammographers reviewing the cases will collaboratively select a single representative screening exam as the index missed case.

3. Case is a minimum of 1 year old (to avoid any interference with clinical care), unless bilateral mastectomy has been performed, or unless films clinically equivalent to those entered into the study from other years are available.

4. Case is not involved in any medical-legal action.

5. No copy films will be used that include significant marks made by a previous observer prior to the copying, and no originals with such permanent marks will be used.

6.1.2 Development of database of missed lesions

The database is nearly complete. All 100 cases with a missed cancer have been identified, although not all have been digitized or categorized. Over half the normals have been collected, with 160 cases from the University of Chicago. The remaining normals will be collected from the University of New Mexico and the University of Chicago. Three tables in the Appendices summarize some of the characteristics of the cancers entered into our database. The average size of the cancers is 11.7 mm.

6.1.3 Computer analysis of case

We will run the computer CAD program on the database, once the database has been completed. This will allow us to use the most current version of our detection schemes. It will take approximately 1 week to run and print the computer results.

6.1.4 Observer study

The formal observer study has not yet begun. We expect to begin this study as soon as all the cases are collected, with a goal of completing the major portion during calendar year 2000.

As a method of testing the design for the observer experiment and collecting observer data, a simulated screeening exercise was conducted, with 100 cases presented to over 100 observers. 50% of these were cancers; none were *missed* lesions. Interesting results were obtained, with a mean sensitivity for general radiologists of 70%, with a broad range from 8% to 98% (standard deviation 14%), and specificity of 68%, with a range of 18 to 91%. The mean sensitivity for 3 experts tested was 81% (range 76% to 86%), with specificity of 54%. Standard deviation of the mean was 3% in both cases, indicating a real difference between experts and general radiologists. While the experts had increased sensitivity, their specificity was lower in this particular test. The details have been published.[10] The results of this were valuable, as a large variation was seen in observer performance, and this will need to be acknowledged and accounted for in the conduct and analysis of the missed lesion/CAD experiment for the remainder of this project. This information is important in designing the conduct of any evaluation of new technology which involves human observers, as the differences between the observers can serve to mask or vitiate otherwise important findings. This has come to light in a practical form with the difficulties encountered by manufacturers of full-field digital mammography devices applying for FDA approval, where large observer variation in the data presented for approval was noted. It is also clear that testing situations do not necessarily mirror routine clinical practice, as all of the cancers in this preliminary study were detected in routine screening, but even experts missed 1 in 5 of the known cancers in this test situation. As a final point, the categories of lesion and reasons for observer oversights resulting from this exercise are reinforcements and guides for the categorization and analysis of the final observer experiment.

6.1.5 Data Analysis

Data analysis of a preliminary observer study[10] has been reported. Data analysis of the missed lesion/CAD study cannot begin until the observer study has been completed.

6.2 Discussion

Progress has been slow over the last year, because Dr. Schmidt, who was the initial PI on the project, transferred from the University of Chicago to New York University. During that time, case accrual has been slow and all other activities were essentially put on hold, with approval from the Army, for about 1 year. Now that Dr. Nishikawa has taken over the role as PI and that Dr. Schmidt has settled in at New York University, we anticipate that we will finish case collection and begin the observer study in the year 2000.

6.3 Recommendations in relation to the Statement of Work

- We do not anticipate making any changes to the Statement of Work.

7. KEY RESEARCH ACCOMPLISHMENTS

- Database is nearly complete with 100% of missed cancer cases and 50% of normal cases collected.
- Characteristics of missed cancers are being compiled.

8. REPORTABLE OUTCOMES

- Since the prior report, a publication[10] has come out from this work, related to testing of radiologists in a simulated screening situation:

  Schmidt RA, Newstead GM, Linver MN, et al., "Mammographic Screening: Sensitivity of General Radiologists," In Karssemeijer N., et al. (eds.), *Digital Mammography - Nijmegen, 1998,* (Kluwer Academic, Dordrecht, The Netherlands, 1998), pp. 383-388

  This article is attached at the end of this report.

9. CONCLUSIONS

Data collection is nearly complete and so we will begin to conduct our main observer study in year 2000. Valuable data has been collected from a preliminary smaller observer study, which will influence the design of the larger scale observer study. We anticipate that we will be able to demonstrate that CAD can reduce the number of missed cancers by 50%, which has not yet been shown in a structured observer experiment. These results should provide information on which health care providers and governmental organizations can base decisions on the value of introducing this promising new technology into the clinical practice of breast cancer screening.

10. REFERENCES

1. C. E. Metz and J. H. Shen, "Gains in accuracy from replicated readings of diagnostic images: prediction and assessment in terms of ROC analysis," Medical Decision Making **12**, 60-75 (1992).

2. E. L. Thurfjell, K. A. Lernevall and A. A. S. Taube, "Benefit of independent double reading in a population-based mammography screening program," Radiology **191**, 241-244 (1994).

3. R. M. L. Warren and S. W. Duffy, "Comparison of single reading with double reading of mammograms, and change in effectiveness with experience," British Journal of Radiology **68**, 958-962 (1995).

4. C. A. Beam, D. C. Sullivan and P. M. Layde, "Effect of human variability on independent double reading in screening mammography," Academic Radiology **3**, 891-897 (1996).

5. J. Brown, S. Bryan and R. Warren, "Mammography screening: An incremental cost effectiveness analysis of double versus single reading of mammograms," British Medical Journal **312**, 809-812 (1996).

6. R. A. Schmidt and R. M. Nishikawa, "Digital screening mammography," Principles & Practice of Oncology **8**, 1-16 (1994).

7. R. A. Schmidt, R. M. Nishikawa, R. Osnis, K. L. Schreibman, M. L. Giger and K. Doi, "Computerized detection of lesions missed by mammography," in *Digital Mammography '96,* edited by K. Doi, M. L. Giger, R. M. Nishikawa and R. A. Schmidt (Elsevier Science, Amsterdam, 1996), pp. 105-110.

8. J. Roehrig, T. Doi, A. Hasegawa, B. Hunt, J. Marshall, H. Romsdahl, A. Schneider, R. Sharbaugh and W. Zhang, "Clinical results with R2 Imagechecker system," in *Digital Mammography - Nijmegen, 1998,* edited by N. Karssemeijer, M. Thijssen, J. Hendriks and L. van Erning (Kluwer Academic Publishers, Dordrecht, The Netherlands, 1998), pp. 395-400.

9. G. M. te Brake, N. Karssemeijer and J. H. C. L. Hendriks, "Automated detection of breast carcinomas not detected in a screening program," Radiology **207**, 465-471 (1998).

10. R.A. Schmidt, G.M. Newstead, M.N. Linver, G.W. Eklund, C.E. Metz, M.A. Winkler, R.M. Nishikawa, "Mammographic Screening: Sensitivity of General Radiologists," in *Digital Mammography - Nijmegen, 1998,* edited by N. Karssemeijer, M. Thijssen, J. Hendriks and L. van Erning (Kluwer Academic Publishers, Dordrecht, The Netherlands, 1998), pp. 383-388.

11. APPENDICES

Table 1. Distribution of breast density in our database

Breast Density	Frequency of Occurrence
Normal	0.30
Fatty	0.21
Dense	0.37
Focal	0.09

Table 2. Distribution of subtlety on a 5-point scale, where 1 is extremely subtle.

Subtlety Rating	Frequency of Occurrence
1	0.16
2	0.39
3	0.37
4	0.05
5	0

Table 3. Distribution by lesion type*

Type of Lesion	Frequency of Occurrence
Asymmetric Density	0.29
Archectural Distortion	0.24
Developing Density	0.07
Mass	0.46
Calcifications	0.10

*numbers sum to greater than 1, because some cases have multiple lesions.

Table 4. Distribution of possible reasons for cancers being missed.*

Possible Reason	Frequency of Occurrence
Seen on only 1 view	0.48
Obscured by overlying tissue	0.40
Looks like normal tissue	0.36
"Busy" breast	0.29
Film technique	0.26
Distracting lesions	0.24
Subtle lesion	0.14
Marginal lesion	0.10
Developing density	0.10
Benign appearing lesion	0.07
Lack of prior films	0.07
Too small to prompt workup	0.05
Lucent lines	0.05
Stable lesion	0.02

*numbers sum to greater than 1, because up to three reasons were given per case.

MEASUREMENT SYSTEM
(measure in mm)

```
MLO view        CC view
 / | \           / | \
(  |  )        (<--+-->)AP
 \_|_/           \_|_/
   CC              ML
```

REVIEW FORM FOR
CANCERS MISSED ON MAMMOGRAPHY
DAMD17-96-1-6229

MAMMOGRAPHY PROJECT 1: MISSED LESIONS DATA ENTRY FORM

COMMENT ______________________________

REVIEWED BY: BS__ KS __ DW__ GB__ JS__ DATE __/__/__ ENTERED TO DB__/__/__

CLINICAL HISTORY AT TIME OF CLINCIAL INDEX CASE

PALPABLE _ PosFamHx ______ *PRE/POST*-MENOPAUSE H/O CANCER ______ YrDx____

PALPABLE	PosFamHx		H/O CANCER
N	*NONE*		*NONE*
Y	*MOTHER*		*BREAST*
?	*SISTER*	*SISTERS*	*OVARY*
	GRANDMOTHER - M/P		*OTHER*

#PrevBx:R _ L _ BASELINE FILM *NORMAL* *ABNORMAL* *NONE* *INADEQUATE* DATE__/__/__

CLINICAL INDEX CASE: THE MAMMOGRAM ON WHICH THE LESION WAS "MISSED"

DATE __/__/__ SIZE:CC____ AP____ ML____ CM POST NIPPLE _____ RAD ___

TYPE __________ SIDE _ QUADRANT __ O'CLOCK ___ BREAST DENSITY _____

TYPE	SIDE	QUADRANT	BREAST DENSITY
MASS	*L*	*UI*	*NORMAL*
MICROCALCS	*R*	*UO*	*DENSE*
MASS+CALCS		*LI*	*FATTY*
ASYMMETRY		*LO*	*FOCAL*
DISTORTION		*RetroAreolar*	
DEVEL.DEN.		*Central*	

SUBTLETY ___ REASON MISSED __ __ __ RECOMMEND ______________

REGARDING THE FIRST MAMMOGRAM ON WHICH THE LESION WAS "FOUND"

DATE __/__/__ SIZE:CC____ AP____ ML____ RAD ___

SUBTLETY ___ REASON FOUND ______________

RECOMMEND ______________

PATHOLOGY ______________

FOLLOW-UP STATUS ______________ FU DATE __/__/__

OTHER RELEVANT MAMMOGRAMS[*B*aseline,*I*ntervening,*X*eromam,*U*ltrasound,*F*ollowup]

C: DATE__/__/__ SIZE:CC____ AP____ ML____ SUBTLETY ____ RELEVANCE *B I X U F*

D: DATE__/__/__ SIZE:CC____ AP____ ML____ SUBTLETY ____ RELEVANCE *B I X U F*

CAD INDEX CASE: DATE __/__/__ COPY? __ *N Y* RESULT ______________

REASONS FOR "MISSING" INDEX CASE

0) Observer Miss
1) Interpretation Error
2) Lesion Not on Film
3) Marginal Lesion
4) Film Technique
5) "Busy" Breast
6) Distracting Lesions
7) Benign Appearing Lesion
8) Developing Density
9) Stable Lesion
10) Too Small to Prompt w/u
11) Looks Like Normal Tissue
12) Lack of Prior Films
13) Lack of comp with Physical Findings
14) Seen on only 1 View
15) Subtle Lesion
16) Lucent Lines
17) Obscured by Overlying Tissue
18) Lack of Follow-up

DEGREE OF SUBTLETY FOR DETECTION

0) Cannot Detect Lesion, even in retrospect
1) Extremely Subtle......(unlikely to be detected by expert mammographer)
2) Very Subtle...........(likely to be detected by expert mammographer)
3) Subtle................(likely detected with good mammographic training)
4) Relatively Obvious....(likely detected w/limited mammographic training)
5) Obvious...............(readily apparent to untrained observer)

Schmidt RA, Newstead GM, Linver MN, et al., "Mammographic Screening: Sensitivity of General Radiologists," In Karssemeijer N., et al. (eds.), *Digital Mammography - Nijmegen, 1998,* (Kluwer Academic, Dordrecht, The Netherlands, 1998), pp. 383-388

MAMMOGRAPHIC SCREENING: SENSITIVITY OF GENERAL RADIOLOGISTS

ROBERT A. SCHMIDT[1], GILLIAN M. NEWSTEAD[2], MICHAEL N. LINVER[3], G.W. EKLUND[4], CHARLES E. METZ[1], MICHAEL A. WINKLER[1], AND ROBERT M. NISHIKAWA[1]

[1] *Kurt Rossmann Laboratories for Radiologic Image Research, Department of Radiology, The University of Chicago, Chicago, IL, USA*
[2] *Dept. of Radiology, New York University, NY, NY, USA*
[3] *Dept. of Radiology, University of New Mexico at Albuquerque, and X-Ray Associates of New Mexico, Albuquerque, NM, USA*
[4] *University of Illinois College of Medicine at Peoria, Peoria, IL, USA*

1. Introduction

High quality mammography can detect early, curable breast cancer and decrease mortality. Much research effort is being expended to improve mammography (digital mammography, computer-aided diagnosis [CAD]), and develop alternative modalities (ultrasound, MRI, radionuclide imaging). However, the human observer is at this point potentially the weakest link in the diagnostic imaging chain, and the range of performance in routine practice is unknown. We have conducted a large observer study using a standardized test base to further investigate this issue.

2. Materials and methods

We did our study at five meeting locations in the US in 1997 and 1998, using films selected from three clinical mammography practices. High quality copy films of 100 cases were presented to a total of over 250 observers who were attending these conferences, and 4 selected experts. Films were displayed (without prior studies or clinical history) on motorized viewboxes designed for mammography, and observers given about 2 1/2 hours to complete the exercise, in supervised workshop settings that typically had 1 to 3 radiologists per viewbox. The case mix was 45 cases containing 50 cancers diagnosed in routine practice, and 55 normal/benign cases. Data were collected regarding the level of experience of observers and the number of mammograms they read. We have graded the first 100 observers for this report.

The distribution of breast lesions in the test set mirrored that in clinical practice, with an emphasis on masses, distortions and asymmetries, rather than calcifications. Microcalcifications probably account for 40 to 50% of tissue sampling breast interventional procedures in the US, but their average positive biopsy yield is lower than that of masses, particularly after the introduction of less invasive percutaneous needle sampling techniques. The perceptual problems in screening associated with detection of significant soft tissue abnormalities is considered harder than the detection of microcalcifications by the authors of this paper, and invasive cancers are more life threatening; hence the emphasis on this type of potentially missed lesion. The distribution of morphologies on mammogram of the breast cancers was: spiculated mass (Mass-S) - 42%, circumscribed mass (Mass-C) - 6%, architectural distortion (ARD) - 14%, asymmetric density (ASD) - 6%, mass + calcifications - 6%, mass + ARD - 6%, ASD or ARD + calcifications - 10%, microcalcifications only (Ca^{++}) - 10%. Figure 1 shows the relative number of lesions of different types, graded by our assessment of their mammographic suspicion (BIRADS-type rating, with 5 being the highest suspicion).

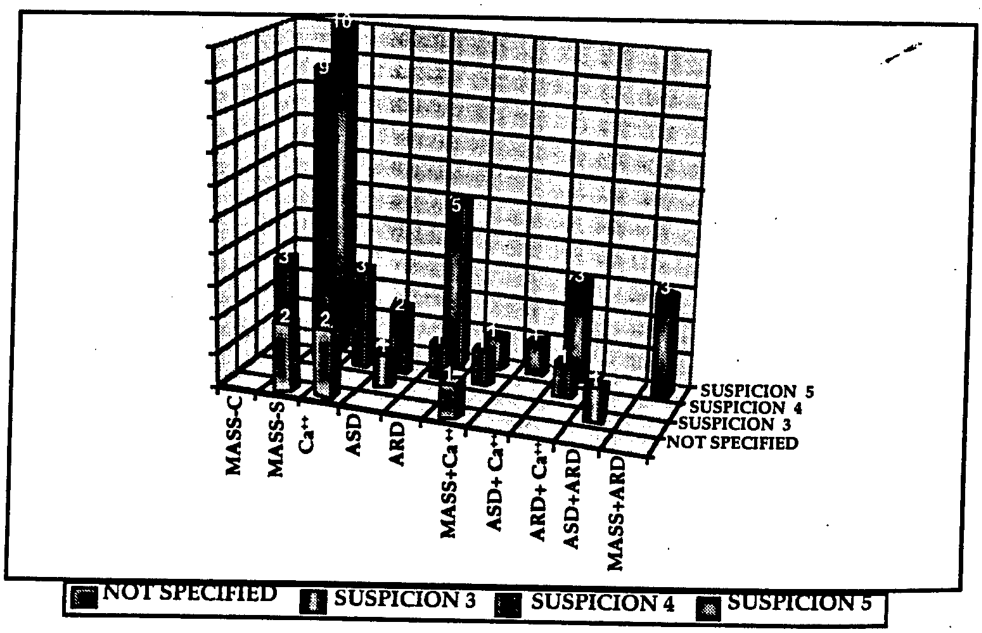

Figure 1. Distribution of lesion types by mammographic suspicion. The terms are defined in the text.

84% of the cancers were invasive, and 16% in situ. The distribution of pathologic types of the breast cancers was: infiltrating ductal carcinoma (IDC) - 48%, IDC + ductal carcinoma in situ (DCIS) - 12%, infiltrating lobular carcinoma (ILC) - 12%, IDC/ILC -

2%, tubular IDC - 6%, medullary IDC 2%, papillary IDC 2%, pure DCIS - 16%. The cancers chosen were representative of average difficulty cases encountered in routine screening practice. They were not the most subtle or tricky cases, and none were cancers which had been "missed," although a typical fraction had films prior to the index (detected) case where the cancer could have been diagnosed. These prior films were not shown to the observers. At the first session only, previous mammograms were hung above the test cases for comparison, but this slowed down the observers, and made it difficult to maintain the pace needed to complete the exercise in a reasonable time.

Mammograms were from examinations taken since 1986, with the majority comparable in technical quality to the range of examinations seen in current clinical practice. The same films were shown to all participants. All except one case (a unilateral examination) had two views of each breast, hung on RADX Mammoscope viewers brought to the meeting, with MLO and CC views hung with right and left views back to back. Light restricting shutters were used, room lights dimmed and magnifying lenses made available, to simulate normal clinical practice Approximately 12 cases were hung on each of 8 automated viewers. A two part NCR carbonless form was devised for scoring (figure 2), so that participants could retain one copy while going over the answers to the cases with an expert at the viewboxes, at the end of the session. This also ensured that answers were not altered at the time of review. Observers were asked to mark whether the case was normal (corresponding to BIRADS codes 1 and 2) or abnormal. If an abnormality was detected, they were asked to mark the lesion type, location on two views, if possible, and their level of suspicion on a five point scale. In subsequent work, we have used a 10 point suspicion scale, to generate ROC curves. Readers were told there were more normal than abnormal cases, and were given about 1 minute per case, with the structured exercise lasting 2 1/2 hours. An additional 1 1/2 hours were devoted to going over the individual cases with participants in small groups.

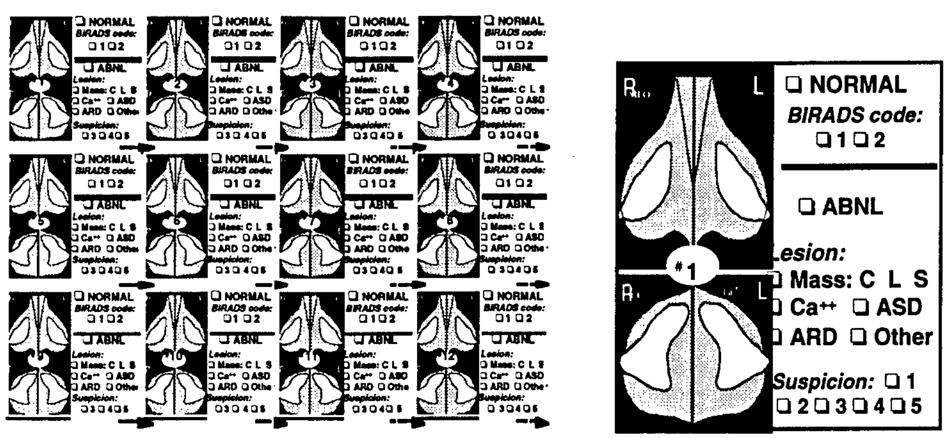

Figure 2. Sample scoring sheet, and enlargement to show detail

If there was more than one lesion, the observers were asked to mark these "#1," "#2," etc. The case set did contain an enriched mixture of bilateral cancers: 5 cases out of 45 patients with cancer, or 11%. Synchronous bilateral cancer is usually only seen in about 3% of screening detected cancers, but we knew from previous work that such cases are disproportionately represented in series of cases that radiologists miss.

Answers were graded correct if the case was marked abnormal and the correct location of the cancer was marked on at least one view, judged as the mark within a distance approximately 1/3 of the distance between the nipple and the chest wall from the true location. If a positive case was left blank, a false negative was scored. If a negative case was left blank, a false positive was scored. Only participants who answered more than 90% of the 105 "cases" (55 normals, and 45 abnormals containing 50 cancers) were considered to have completed the exercise. This eliminated 25% of the 100 readers, and 1 of the 4 experts, who were thereby dubbed "non-compliers." Results are given in Figure 3 for those who completed the test, and in Figure 4 separately for the 22 non-complying radiologists (and 3 non-radiologists). For the latter figure, however, only the cases for which an answer was given were graded. On average, experts, complying radiologists and non-compliers answered 99.7%, 97.3% and 79.1% of the cases, respectively.

3. Results

Based on analysis of the first 100 observers and four experts, there is a considerable range of accuracy in reading screening mammography among general radiologists in the US, and expert mammographers are generally better at the screening task (Figure 3):

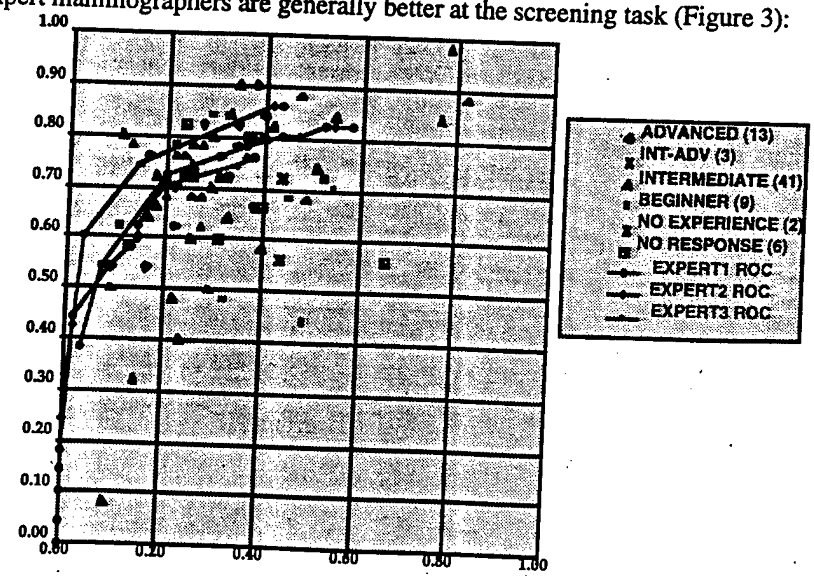

Figure 3. Scatter plot of true positive fraction [TPF, ordinate] versus false positive fraction [FPF, abscissa] for 75 general radiologists, and ROC curves for 3 experts

For the observers who gave answers to more than 90% of the cases, the average sensitivity was 70%, with range of 8% to 98% (SD 14%; spec = 68%, with range of 18 to 91%) for correct cancer detection and localization for 75 general radiologists, and 81% for 3 experts, with smaller range (76% to 86%; specificity = 54%). Standard error of the mean was 3% in both cases. There was only a relatively weak correlation with general observers' self-assessment of their level of expertise (Figure 5). The sensitivity of those who did not complete 90% of the cases was only 42%, with specificity of 72%.

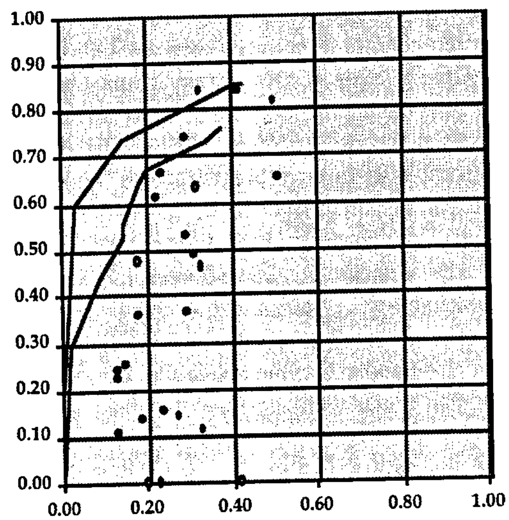

Figure 4. Scatter plot of TPF versus FPF for the 25 "non-complying" radiologists, and the ROC curves of the best and worst of 3 experts

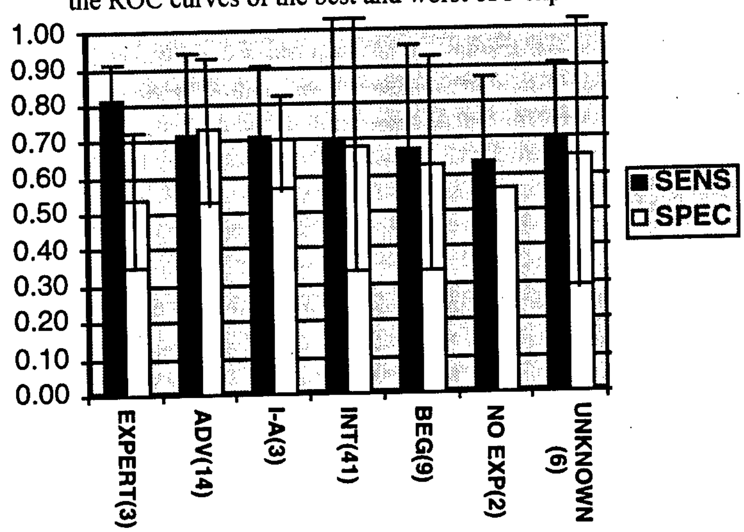

Figure 4. Sensitivity and specificity, based on radiologists' self assessment. Experts were designated by the authors. Error bars are two standard deviations.

4. Discussion

This exercise was modeled after the second level Teaching Course in Mammography of László Tabár. We individually scored the participants (generating 30,000 data points), and they were given the option of getting specific quantitative feedback on their performance; more than half took this option. A certificate for reading 100 cases under supervision can be given towards MQSA requirements in the US.

While the results may not be unexpected, the range of performance in detecting breast cancers on screening mammography by general radiologists is quite large, and radiologists who are experts and dedicated to mammography perform substantially better than the average radiologists, detecting about 16% more cancers in this study. This increased sensitivity comes at the price of decreased specificity, however. This increase in detection rate is comparable to improvements that can be confidently expected from improvements in the mammogram images themselves, or by developing alternative modalities. The introduction of computer-aided diagnosis (CAD) techniques into clinical practice would be expected to decrease the gap between the average reader and the expert reader, and decrease the variability of readings, but this will require further large scale studies. There are also obvious implications for improving the training of radiologists, and establishing competency standards, which have not yet been implemented in the US.

5. Conclusions

General radiologists read mammograms with higher specificity and lower sensitivity than experts. There is room for improvement in breast cancer detection: experts are at least 16% more sensitive than general radiologists, and the variability of general radiologists is very high. There is a need for improved training and feedback for radiologists, with indication of a need for minimum competency testing. Benefits similar to those expected from imaging technology advances are likely possible, and one way that performance may be improved through technical advances is by use of CAD.

Acknowledgments

This work was supported in part by a grant from the US Army (DAMD17-96-6229). The contents of this paper are solely the responsibility of the authors and do not necessarily represent the official views of any of the supporting organizations. The authors gratefully acknowledge the assistance of David Joyner of RADX corporation, for providing us with the mammography viewstations used at each of the meeting locations, and of Brian Goode of Faxitron, in providing magnifying lenses for the participants for the last several sessions.